Semantic-Guided Radar-Vision Fusion for Depth Estimation and Object Detection

Wei-Yu Lee
WeiYu.Lee@UGent.be
Ljubomir Jovanov
Ljubomir.Jovanov@UGent.be
Wilfried Philips
Wilfried.Philips@UGent.be

TELIN-IPI
Ghent University-imec
Gent, Belgium

Abstract

In the last decade, radar is gaining its importance in perception modules of cars and infrastructure, due to its robustness against various weather and light conditions. Although radar has numerous advantages, the properties of its output signal also make the development of fusion scheme a challenging task. Most of the prior work does not exploit full potential of fusion due to the abstraction, sparsity and low quality of radar data.

In this paper, we propose a novel fusion scheme to overcome this limitation by introducing semantic understanding to assist the fusion process. The sparse radar point-cloud and vision data is transformed to robust and reliable depth maps and fused in a multi-scale detection network for further exploiting the complementary information. In our experiments, we evaluate the proposed fusion scheme on both depth estimation and 2D object detection problems. Object detection results compare favourably to the state-of-the-art and demonstrate the effectiveness of the proposed scheme. Depth map estimation results are on par with the state-of-the-art on depth from RGB estimation. The ablation studies also show the effectiveness of the proposed components.

1 Introduction

With the rapid need to develop the autonomous driving system, object detection and depth estimation have become increasingly important research problems. For achieving accurate and reliable results, multiple sensors are fused to take advantage of different modalities, especially LiDAR, radar, and camera. Radar is an important sensor in perception modules because of its robustness against various weather and light conditions, but it has not been investigated much in a fusion context. Moreover, compared to LiDAR, radar has a larger operation range and lower costs [1, 16], which makes it a suitable alternative for combining with RGB cameras. However, due to the specular nature of electromagnetic reflections at wavelengths employed by the radar [30], noisy measurements with a limited field of view make designing an applicable fusion scheme become a challenging problem with limited previous research.

In order to fuse the radar signal with RGB images, numerous methods have been proposed to tackle this problem. Typically, radar data is processed by CFAR algorithm to convert the range-azimuth-doppler tensor into a point-cloud to separate the targets of interest. There are a few studies combining the radar point-cloud with various fusion schemes, was presented in [8, 20, 21, 22, 23, 30]. Although the surrounding clutter and noise have been eliminated, the measurements are still ambiguous and abstract in terms of position and velocity. Furthermore, the fusion is also limited due to the extremely sparse points (much fewer than LiDAR). Hence, it becomes challenging to fully exploit the complementary information of the radar and RGB images.

In this paper, we propose a novel scheme for 2D object detection and depth estimation based on semantic-guided fusion which combines the radar point-cloud and vision data. We are aiming at fusion of the radar point-cloud with monocular RGB images on two levels. Firstly, in order to alleviate problems caused by the sparsity and abstraction of the radar point-cloud, we introduce a semantic-guided depth estimation method to fuse vision data and generate accurate dense depth maps. We rely on the semantic understanding of vision data and the long detection range of radar signals to improve the performance of depth estimation and enhance the visibility of far targets in the scene. With the proposed method, the sparse radar point-cloud could be well transformed into a robust and reliable depth map, solving the limitation introduced by radar signal post processing. One of the advantages radar brings into this fusion, over RGB based depth estimation methods is physical distance of the targets.

Moreover, in the second level, we further combine the two sensor modalities to perform object detection by proposing a multi-scale fusion network (FusionYOLO) based on YOLOv3 [25]. For fully exploiting the potential of the complementary information from the previous level, an additional extraction network is introduced to explore the features from modalities. We fuse the estimated depth maps with vision data, relying on multi-scale representations. Figure 1 illustrates the main architecture of our work. The two levels are trained separately from scratch and fine-tuned together with the object detection goal.

Our contributions can be summarized as follows:

- To our best knowledge, we are among the first to propose a novel semantic-guided fusion scheme, integrating features from monocular RGB images, semantic information, and sparse radar point-clouds to estimate accurate dense depth maps and to perform 2D object detection.
- Our proposed multi-scale fusion framework (FusionYOLO) and estimated dense depth maps can significantly enhance the visibility of small and far targets in the scene.
- In our experiments, we qualitatively and quantitatively verify the performance of our model and achieve the state-of-the-art 2D object detection performance and a comparable depth estimation result on *nuScenes* dataset.

2 Related Work

2.1 Depth Estimation

RGB-based Depth Estimation One of the main research directions of RGB-based depth estimation is relying on the supervision from stereo image pairs or video sequences. Various objective functions were proposed to explore the geometric priors, such as inverse warping

[6], left-right viewpoints consistency [7], or spatial and temporal warping [31]. On the other hand, without using additional sources of information, estimating depth solely from monocular images is much more challenging. Eigen et al. [5] proposed a multi-scale framework for estimation of the depth map, relying on the coarse and fine cues. Liu et al. [18] combined the strength of deep convolution network and continuous condition random field to estimate the depth. In addition to exploiting the geometric properties from image pairs or the depth ground truth, some prior works also introduced the semantic information to assist the depth estimation. Most previous methods used shared latent representations [12], task-specific sub-network [4, 32], or multi-task learning to leverage the semantic information [9].

Cross-modal Depth Estimation In contrast to monocular depth estimation, combining information from multiple modalities, results in more reliable and accurate results. LiDAR is one of the most popular sensors used for fusion with RGB data. Jartiz et al. [11] proposed a late fusion scheme to handle LiDAR data and RGB images to learn semantic segmentation and depth completion. Ma et al. [19] simply concatenated the depth information with vision data along RGB channels to learn directly from RGB-D data. Qiu et al. [24] used the estimated surface normal as the intermediate representation to produce dense depth. Nevertheless, due to the sparsity, noisiness and lower quality of radar point-clouds, it is hard to apply these existing LiDAR fusion methods on radar and vision fusion problems [16].

On the other hand, the literature on using radar point-clouds and RGB images for dense depth estimation is quite limited. Lin et al. [16] proposed a two-stage framework for reducing noise in radar measurements and dense depth estimation. They used the coarse estimation to learn a filtering process, which removes the outliers from the radar data in the first stage, then a similar network is applied in the second stage for refinement. However, due to the lack of depth ground truth in the dataset, LiDAR data is used to supervise the training. Although LiDAR data contains more accurate and dense depth measurements, its shorter sensing range could easily remove a significant amount of information in the radar signal, and limit the valid detection range of the estimated depth maps.

2.2 Radar-Vision Fusion-based Object Detection

Another important line of work related to our method is radar-vision fusion for object detection. Most of the previous works directly utilizes the sparse radar point-cloud as complementary information to the existing detection networks such as Faster R-CNN [26]. Nabati et al. [20] introduced a radar-based region proposal network with pre-defined anchor boxes to improve the detection accuracy. Furthermore, in [21], a middle-fusion approach was proposed to generate object proposals by utilizing both radar and image features. Nobis et al. [23] simply concatenated and fed the sparse radar data into a pre-trained VGG to learn which level is most beneficial for object detection. Chadwick et al. [2] also proposed an additional network to extract and concatenate the radar features to an image-based model for detecting small objects. Yadav et al. [29] used an attentive feature pyramid network to highlight the important radar features and fused with RGB images. Chang et al. [3] also generated an attention matrix to control the information flow within vision sensor. These existing research demonstrate that radar and vision data are quite complementary. However, the sparsity and noisiness of radar measurements still limits the fusion development. We can conclude from the above that direct fusion of the noisy and abstract radar point-cloud with the dense RGB images fails to fully exploit the complementary information of the sensors [30].

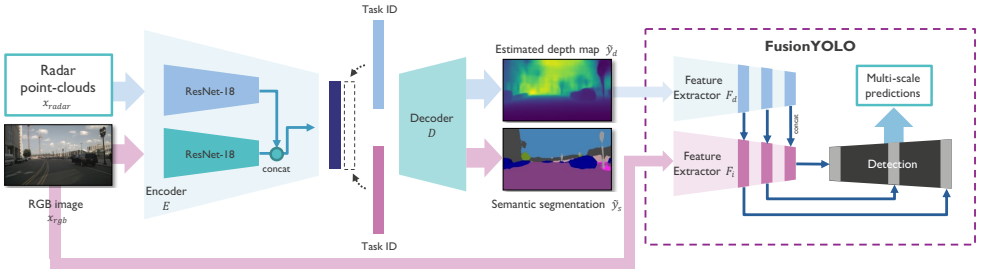

Figure 1: Our proposed semantic-guided fusion framework. With jointly learning semantic segmentation and depth estimation, the sparse and noisy radar point-cloud could be transformed into a robust and reliable depth map, solving the limitation introduced by radar post processing and allowing the model to fully exploit the information of different modalities.

3 Proposed Method

The goal of our proposed fusion scheme is to perform object detection and depth estimation using both radar point-clouds and vision data as input. Our model can be separated into two parts: cross-modal depth estimation and object detection framework. As illustrated in Figure 1, during training, we use RGB images x_{rgb} and the projected 2D radar point-cloud x_{radar} as inputs. The outputs from the proposed network are: the estimated depth map $\tilde{y}_d$ and semantic segmentation $\tilde{y}_s$. The first level model jointly learns depth and semantic information by using a shared decoder, which is trained using LiDAR measurements y_d and guided by the semantic segmentation labels y_s simultaneously. In the second level, our cross-modal detection network uses the intermediate depth maps $\tilde{y}_d$ and vision data x_{rgb} as inputs, and outputs the detection results. We extract the features by two independent feature extractors F_i and F_d. Finally, we merge the different modalities on multiple scales for considering meaningful semantic and depth information from coarse to fine-grained features.

3.1 Cross-modal Depth Estimation

Depth Estimation In previous research, LiDAR measurements are typically treated as the ground truth to supervise the depth estimation [16]; however, due to the shorter sensing range of LiDAR, far targets detected by the radar in the scene would be easily ignored. Hence, in order to preserve the long distance sensing information from radar data and estimate an accurate depth map covering the maximal sensing range of radar, **during model training** we propose to use LiDAR measurements and semantic information from RGB images instead of using LiDAR as the only ground truth. Specifically, because the semantic segmentation of the scene can recognize the far objects, we aim at leveraging the semantic information to give the model a hint to estimate the depth of the far targets.

Inspired by [4], we use a shared decoder to learn depth estimation and semantic segmentation concurrently with different task identity t (as shown in Figure 1) to control the objective of training: $f_s(x_{rgb}, x_{radar}, t) = D\left(E\left(x_{rgb}, x_{radar}\right), t\right)$, where D, E are the decoder and encoder. The task identity t is appended to the extracted feature maps from the encoder. According to the different assigned task identity t, we compute the depth estimation loss $\mathcal{L}_{depth}$ by assigning $t = 1$ with element-wise L1 loss: $\mathcal{L}_{depth} = \left\| f_s(x_{rgb}, x_{radar}, 1) - y_d \right\|$, and compute the semantic segmentation loss $\mathcal{L}_{seg}$ by assigning $t = 0$ with cross-entropy CE:

$\mathcal{L}_{seg} = CE\left(f_s(x_{rgb}, x_{radar}, 0), y_s\right)$. Thanks to this unified decoder, the two training objectives share the same parameters and transfer the geometric projection and semantic information across the different modalities.

Semantic Guidance For reinforcing the semantic awareness when estimating depth of long distance targets, we further introduce a guidance loss function $\mathcal{L}_{gd}$. In order to encourage the depth map to preserve the far and small targets and to be locally smooth, we do not only consider the original RGB images [7, 8, 16] but also the semantic segmentation results concurrently to develop a regularization term:

$$\mathcal{L}_{gd} = \left(\left|\partial_x \tilde{y}_d\right| e^{-\left|\partial_x x_{rgb}\right|} + \left|\partial_y \tilde{y}_d\right| e^{-\left|\partial_y x_{rgb}\right|}\right) + \phi\left(\tilde{y}_s\right) \odot \left(\left|\partial_x \tilde{y}_d\right| e^{-\left|\partial_x \tilde{y}_s\right|} + \left|\partial_y \tilde{y}_d\right| e^{-\left|\partial_y \tilde{y}_s\right|}\right) \quad (1)$$

where $\odot$ denotes element-wise multiplication, and $\partial_x(\cdot)$ and $\partial_y(\cdot)$ indicate the gradients in two different directions. Different from other methods [7, 8, 16], we additionally involve the edge-aware constrains from the semantic segmentation. However, due to the semantic segmentation output $\tilde{y}_s$ being also learned by the model during the training, unified weightings between the RGB images and semantic information constrains might mislead the network to over-smooth the regions with less confidence. Hence, for proper weighting the semantic awareness cost, we propose a confidence map $\phi(\tilde{y}_s)$ to reallocate the guidance priority, where $\phi(\tilde{y}_s)$ is composed of the highest confidence values at each location of the segmentation map $\tilde{y}_s$ predicted by the model. Specifically, if the model can recognize the objects with high confidence, the regularization term will put more emphasize on the constraints from the recognized objects in the semantic segmentation outputs, and vice versa.

During training, the depth estimation model is trained independently from scratch using the following objective function:

$$\mathcal{L}_{ds} = \mathcal{L}_{depth} + \alpha_{seg}\mathcal{L}_{seg} + \alpha_{gd}\mathcal{L}_{gd} \quad (2)$$

where α_{seg} and α_{gd} are the weights for each loss. During the inference, the model takes radar point-clouds and monocular images to produce dense depth maps or semantic segmentation maps by manipulating the task identity manually.

3.2 Cross-modal Object Detection Framework

Multi-scale Fusion In the second part, we further combine monocular RGB images and radar point-clouds in a multi-scale scheme, called FusionYOLO. Since the modalities have different physical properties and meanings of a single pixel (RGB values and projected 2D radar point-clouds), it is not straightforward to train the network by directly combining information from different modalities. Instead of directly applying the projected 2D radar point-clouds to the network with vision data [2, 3, 23, 29], we propose to use the estimated dense depth maps and RGB images as inputs to perform object detection for fully exploiting the potential of the complementary information from the previous level. Similar to the original feature extractor F_i for RGB images, we introduce an additional extraction network F_d for learning the depth information as illustrated in Figure 1. In order to stabilize the training process and prevent over-fitting problem, we propose to use independent model parameters to build the feature extractors F_i and F_d, still relying on the same network architecture.

Furthermore, for leveraging the features from low to high levels, we follow YOLOv3 [25] to extract the features from depth maps at three different scales and fuse with image

features. Different from the other methods [2, 3, 23, 29], we propose to use dense depth maps from the first part providing more details of the scenes than the sparse and noisy radar point-cloud data to assist FusionYOLO to recognize the objects. In addition, we propose to perform the training of depth estimation and detection models simultaneously. Specifically, the depth maps used by the detection model are also improved by considering the object location and classification labels. Moreover, several possible fusion methods are conducted to compare their performance. According to the experiments, feature concatenation has best performance. Therefore, we adopt concatenation at different scales as our network design.

Learning of the Proposed Framework During training, we separately train the depth estimation and object detection models from scratch and fine-tune the whole network end-to-end. However, finding the optimal weightings of the models with numerous tasks is time-consuming and difficult. Therefore, we balance the two models by learning the weightings as [13]:

$$\mathcal{L}_{total} = e^{-w_1}\mathcal{L}_{ds} + e^{-w_2}\mathcal{L}_{det} + \sum_i w_i \tag{3}$$

where $\mathcal{L}_{det}$ is the objective loss of the detection network, which contains bounding box regression, classification and confidence losses, and w_1, w_2 are learned automatically by the network to balance the two models for achieving higher performance.

4 Experiments

4.1 Dataset and Implementation Details

Dataset In this paper, we use the newly released nuScenes dataset [1] to evaluate our performance. In this dataset, 6 cameras, 5 FMCW radars and 1 LiDAR were used, covering nearly 360 degree of view. The FMCW radar used in this dataset can detect up to 250m, and the LiDAR only can reach 70m from the ego-vehicle. This dataset contains 1000 road scenes recorded in different conditions. For each scene, they synchronized the sensors and released the official tool for sensor coordinate calibration as well as 2D/3D annotation conversion. Annotations of 850 scenes are freely available. In all of our experiments, we rely on official codes to split the data into 700 scenes for training and 150 scenes for testing and project the radar point-clouds to 2D image. In addition, we use samples from the front and rear view of cameras and all the radars for training and evaluation. Furthermore, there are 23 different classes in the dataset, and we follow settings from [20, 21] to condense them into 6 classes: Car, Truck, Person, Bus, Bicycle, and Motorcycle. We also follow [16] to use LiDAR measurements as the ground truth. In addition, because nuScenes dataset [1] did not release the semantic segmentation annotations of the vision data, we introduce the state-of-the-art model [27] to build the pseudo-labels for training.

Implementation In the depth estimation model, for fairly comparing with the reference method of [16], we follow most of their settings and use their first stage network architecture, which is composed of the standard ResNet18 [10] and UpProj [14], to develop our model (only half size of model parameters [16]). We append an additional task identity t during training to the feature maps from the encoder to control the outputs. In the object detection model, we follow the original architecture of the feature extractor to build an additional sub-network F_d for learning the depth information. All the models are trained using a batch size

Table 1: Depth estimation performance comparison. It is worth noting that our model can achieve comparable results by only using half size of the state-of-the-art model parameters [16] by leveraging the semantic information of vision data.

Methods	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	RMSE $\downarrow$	MAE $\downarrow$	REL $\downarrow$	$MAE_{log} \downarrow$
BTS [15]	0.872	0.948	0.976	5.561	2.391	0.123	0.048
Sparse-to-dense [19]	0.876	0.949	0.974	5.628	2.374	0.115	0.047
PnP [28]	0.863	0.948	0.976	5.578	2.496	0.128	0.050
Lin et al. (single-stage) [16]	0.884	0.953	0.977	5.409	2.270	0.112	0.045
Lin et al. (two-stage) [16]	**0.901**	**0.958**	**0.978**	**5.180**	**2.061**	**0.100**	**0.040**
Ours (RGB only)	0.870	0.948	0.976	5.597	2.499	0.126	0.050
Ours	0.895	**0.958**	**0.978**	5.209	2.104	0.104	**0.040**

of 16 and the SGD optimizer with a learning rate of 0.001 and a momentum of 0.9 for 100 epochs. Please refer to Supplementary Materials for details of our network architecture.

4.2 Quantitative Results

Depth Estimation In order to demonstrate the effectiveness of our depth estimation model, we compare the performance with the monocular depth estimation method [15] and other existing fusion methods [16, 19, 28]. The results are shown in Table 1. We adopt mostly used metrics from previous works to evaluate our performance [4, 5, 19]. For fair comparison, we follow [16] to use the same settings and codes as our baseline to develop our model. Also, because of the lack of accurate scene measurements, we follow [16] to use the LiDAR measurements as the ground truth. A fair and comprehensive comparison should also consider far targets ($> 70m$) in the scene. However, LiDAR has only about $70m$ sensing range, which is much shorter than radar ($\sim 250m$). Hence, considering the qualitative results shown in Section 4.3 is necessary.

As we can observe in Table 1, the LiDAR + RGB fusion methods [19, 28] could not achieve higher performance when we simply replaced LiDAR data with the sparse radar point-cloud. Furthermore, we also compare our performance with the state-of-the-art method from [16]. It is worth noting that we only use half size of their two-stage model parameters to achieve comparable results. Specifically, since we leverage the semantic information of vision data, our model can estimate accurate depth map with less model parameters. In addition, our method can also preserve the far radar detection to combine with semantic awareness from vision data to estimate the dense depth map with enhanced visibility of far targets. This enhanced information can assist our next level object detection model to detect far targets. Further details are shown in the next section.

Object Detection We further compare the performance of object detection with the RGB baseline and other fusion methods [20, 21]. We leverage the standard COCO [17] metrics to quantitatively evaluate the performance of the proposed method. AP_{50} and AP_{75} denote the threshold of IoU as 0.50 and 0.75, and AP denotes mean average precision where the threshold of IoU takes values from 0.50 to 0.95, with the step size of 0.05.

As we can find in Table 2, the two-stage baseline Faster-RCNN [26] can achieve higher performance than the single-stage YOLOv3 [25], especially on the strict AP_{75} performance. The trend is similar to COCO dataset [17] results shown in [25]. Due to the different attributes of the single- and two-stage models, such as the network architecture, it is hard to fairly compare with each other but we still list their performance as references. When

Table 2: Quantitative results of object detection on the nuScenes dataset. We also replace the depth maps produced by the monocular depth estimation method [15] and the fusion method [16] to train YOLOv3 [25] and our FusionYOLO to demonstrate the effectiveness of our estimated depth maps.

Methods	Input data	AP	AP_{50}	AP_{75}
Two-stage object detection methods				
Faster R-CNN [26]	RGB	34.95	58.23	36.89
RRPN [20]	RGB + Radar	35.45	59.00	37.00
Nabati et al. [21]	RGB + Radar	35.60	60.53	**37.38**
One-stage object detection methods				
YOLOv3 [25]	RGB	34.69	64.31	33.33
YOLOv3 [25] w/ [15] depth maps	RGB	33.10	63.12	32.08
FusionYOLO w/ [15] depth maps	RGB	34.94	64.82	35.53
FusionYOLO w/ [16] depth maps	RGB + Radar	35.54	65.12	35.83
FusionYOLO w/ our depth maps	RGB + Radar	**35.95**	**65.77**	35.84

Table 3: Object detection results comparison for different sizes on the nuScenes dataset.

Methods	Input data	AP_{Small}	AP_{Medium}	AP_{Large}
YOLOv3 [25]	RGB	25.10	27.61	38.31
FusionYOLO w/ [16] depth maps	RGB + Radar	26.60	28.21	38.43
FusionYOLO w/ our depth maps	RGB + Radar	**27.00**	**28.85**	**38.68**

introducing the radar data, we observe that our method outperforms other two-stage-based methods [20, 21] in terms of AP and AP_{50}. However, although our proposed method do improve the detection performance, we still behind the two-stage based methods [20, 21] on the strict AP_{75} performance. We attribute the performance gap to the different types of baseline models. In addition, in order to demonstrate the effectiveness of our estimated depth maps, we replace the depth maps produced by the monocular depth estimation method [15] and the fusion method [16] to train YOLOv3 [25] and our FusionYOLO. The results show that our depth maps successfully assist our object detection network and achieve significantly better performance than the others.

Moreover, we have also conducted an additional experiment to demonstrate the capability of detection among different target sizes. For fair comparison, we only compare the performance with the single-stage-based methods, and the definition of target sizes are the same as in the COCO dataset [17]. In Table 3, we can see that our proposed model achieves higher performance, especially on small and medium targets. Specifically, although it is hard to separate the targets by their absolute distance from the sensors, we still can observe the similar trends in the size of bounding boxes. The results show that our depth estimation method successfully enhances the visibility of far (small) targets and successfully assists our detection model (FusionYOLO) in object detection. We have also conducted further experiments, such as nighttime/daytime performance, different extraction backbones comparison, and per-class performance analysis. The results are provided in Supplementary Materials.

4.3 Qualitative Results

In Figure 2, we provide the qualitative results of our method, and also show the results of [16, 25] for comparison. One can easily notice that our model produces excellent results with clear edges and less false positives. Especially for the far region in the scene, false

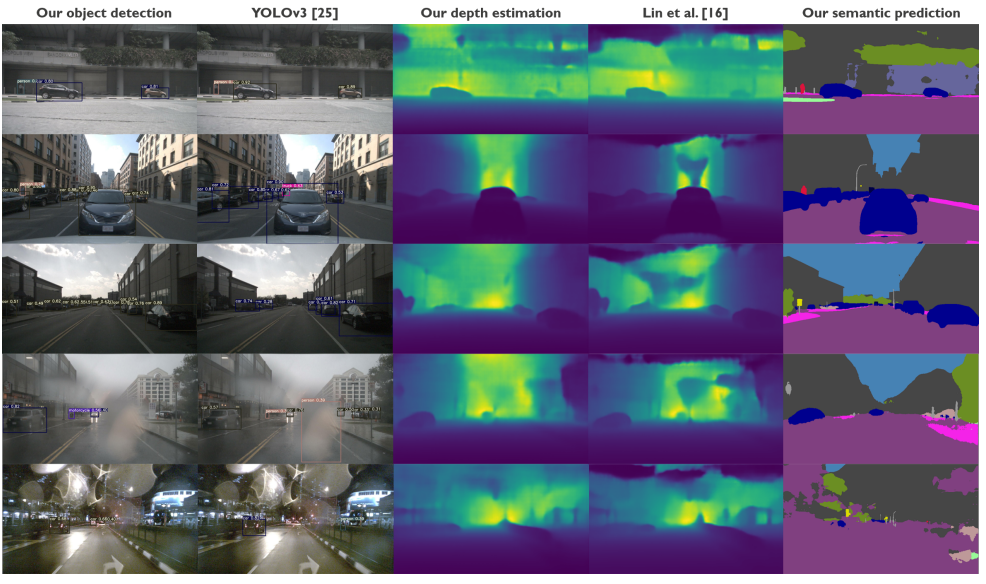

Figure 2: Example results comparisons. Our model performs favorable results with clear edges and less false positives, especially in far regions of the scenes. Even when occlusions cause semantic information loss, our method still estimates depth maps of decent quality.

Table 4: Ablation study of our proposed method on object detection on the nuScenes dataset. The baseline model is our proposed network architecture without radar data and learning semantic segmentation. We evaluate the model with or without our proposed components to verify the improvements.

Methods	Input data	$\mathcal{L}_{seg}$	$\mathcal{L}_{gd}$	$\phi(\tilde{y}_s)$	AP	AP_{50}	AP_{75}
YOLOv3 [25]	RGB				34.69	64.31	33.33
Ours - Baseline	RGB				34.71	64.34	33.41
Ours	RGB	✓			34.94	64.57	33.69
	RGB + Radar				34.97	64.87	34.51
	RGB + Radar	✓			35.47	65.21	35.38
	RGB + Radar	✓	✓		35.87	65.64	35.67
	RGB + Radar	✓	✓	✓	**35.95**	**65.77**	**35.84**

positives of depth values might mislead the object detection model to focus on the wrong regions and also make the inconsistent edges with the vision data. Compared to [16], our method provides a smoother surface without ambiguous predictions. In addition, it is worth noting that even when the occlusions or noise of RGB images cause the loss of the semantic information (see the last row in Figure 2), our method still can estimate depth maps of decent quality. It could be attributed to the proposed confidence map $\phi(\tilde{y}_s)$ in the equation 1. The low confidence scores would suppress the smoothing strength from the semantic awareness, allowing the model to pay more attention on the supervised depth estimation loss $\mathcal{L}_{depth}$. However, we also notice that if the objects are not recognized by both the depth estimation and semantic prediction simultaneously, it might be ignored by the FusionYOLO easily, even the RGB camera successfully captured the objects. (Please refer to the 4th row of Figure 2)

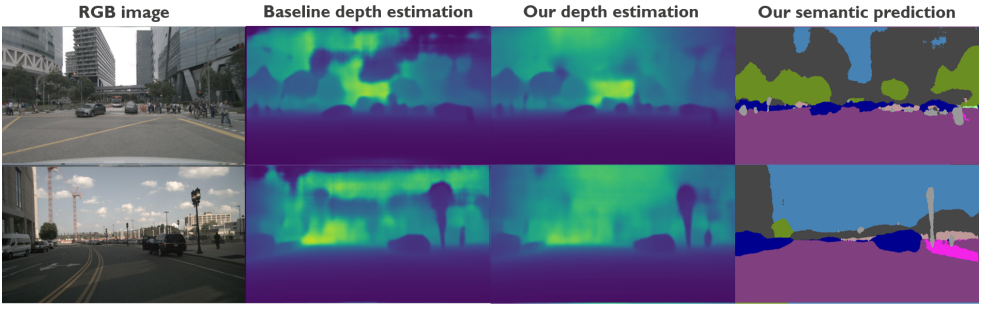

Figure 3: Ablation study of our proposed method on depth estimation.

4.4 Ablation Study

In order to demonstrate the impact of our proposed methods, we conduct ablation studies and show the results in Table 4 and Figure 3. Our contribution of multi-task learning and semantic guidance is evaluated by showing the qualitative results and AP. We use the depth estimation model without using radar point-cloud data and learning semantic information as the baseline, and apply the supervised semantic segmentation loss $\mathcal{L}_{seg}$, semantic guidance loss $\mathcal{L}_{gd}$, and the confidence map $\phi(\tilde{y}_s)$ to show the differences.

First, we evaluate the contribution of semantic segmentation without radar data. In Table 4, considering that we do not introduce any other supervision method, such as left-right viewpoint consistency [4], there is only a marginal improvement when semantic segmentation loss $\mathcal{L}_{seg}$ is applied. When the radar data and semantic information are both taken into account, our model can make a greater improvement. Next, we evaluate the performance with and without semantic guidance loss $\mathcal{L}_{gd}$ and the confidence map $\phi(\tilde{y}_s)$ to verify the effectiveness. Both of them successfully improve the performance by taking advantage of the semantic segmentation map. Finally, the model with all the proposed components systematically outperforms the baseline model with significant improvement on the detection performance and the qualitative results. Our proposed method not only suppresses the false positives in the depth maps, but also enhances the visibility of objects. Moreover, the detection performance also verifies the improvement of the depth maps. The model with all the components achieves the best performance among all the combinations.

5 Conclusion

In this paper, we proposed a novel semantic-guided fusion scheme to combine the radar and vision data to perform depth estimation and 2D object detection. We rely on semantic understanding of vision data to alleviate the limitations of the radar point-cloud, and generate the accurate dense depth maps to enhance the visibility of far targets in the scene, with our proposed semantic guidance loss and unified multi-task architecture. In addition, we combined the vision and estimated depth information in a multi-scale scheme to perform object detection. Quantitative and qualitative results on the public nuScenes dataset confirm that our depth estimation results are on par with the state-of-the-art method, and object detection results compare favourably to the baseline and other models. Our ablation studies also clarify the effectiveness of the proposed components.

Acknowledgements

This work was funded by EU Horizon 2020 ECSEL JU research and innovation programme under grant agreement 876487 (NextPerception) and Marie Sklodowska-Curie grant agreement No. 765866 - ACHIEVE.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] Simon Chadwick, Will Maddern, and Paul Newman. Distant vehicle detection using radar and vision. In *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[3] Shuo Chang, Yifan Zhang, Fan Zhang, Xiaotong Zhao, Sai Huang, Zhiyong Feng, and Zhiqing Wei. Spatial attention fusion for obstacle detection using mmwave radar and vision sensor. *Sensors*, 2020.

[4] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. 2014.

[6] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

[7] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[9] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Maximilian Jaritz, Raoul De Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. In *2018 International Conference on 3D Vision (3DV)*, 2018.

[12] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[13] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, 2016.

[15] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[16] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[18] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 2015.

[19] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[20] Ramin Nabati and Hairong Qi. Rrpn: Radar region proposal network for object detection in autonomous vehicles. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.

[21] Ramin Nabati and Hairong Qi. Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles. In *2020 International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[22] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[23] Felix Nobis, Maximilian Geisslinger, Markus Weber, Johannes Betz, and Markus Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, 2019.

[24] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[27] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[28] Tsun-Hsuan Wang, Fu-En Wang, Juan-Ting Lin, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Plug-and-play: Improve depth estimation via sparse data propagation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[29] Ritu Yadav, Axel Vierling, and Karsten Berns. Radar+ rgb attentive fusion for robust object detection in autonomous vehicles. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2020.

[30] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. Radarnet: Exploiting radar for robust perception of dynamic objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[31] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.